J Autism Dev Disord (2011) 41:311–319
DOI 10.1007/s10803-010-1057-0

ORIGINAL PAPER

Standardized ADOS Scores: Measuring Severity of Autism Spectrum Disorders in a Dutch Sample

Annelies de Bildt · Iris J. Oosterling · Natasja D. J. van Lang · Sjoerd Sytema · Ruud B. Minderaa · Herman van Engeland · Sascha Roos · Jan K. Buitelaar · Rutger-Jan van der Gaag · Maretha V. de Jonge

Published online: 9 July 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract The validity of the calibrated severity scores on the ADOS as reported by Gotham et al. (J Autism Dev Disord 39: 693–705, 2009), was investigated in an independent sample of 1248 Dutch children with 1455 ADOS administrations (modules 1, 2 and 3). The greater comparability between ADOS administrations at different times, ages and in different modules, as reached by Gotham et al. with the calibrated severity measures, seems to be corroborated by the current study for module 1 and to a lesser extent for module 3. For module 2, the calibrated severity scores need to be further investigated within a sample that resembles Gotham's sample in age and level of verbal functioning.

Keywords ADOS · Autism · ASD · Diagnosis · Symptoms · Severity

A. de Bildt (✉) · R. B. Minderaa
Department of Child and Adolescent Psychiatry, University Medical Center Groningen, Accare Groningen, PO Box 660, 9700 AR Groningen, The Netherlands
e-mail: a.de.bildt@accare.nl

I. J. Oosterling · S. Roos · J. K. Buitelaar · R.-J. van der Gaag
Karakter Child and Adolescent Psychiatry University Center, Nijmegen, The Netherlands

N. D. J. van Lang
Department of Child and Adolescent Psychiatry, Leiden University Medical Center, Curium, Leiden, The Netherlands

S. Sytema
Department of Psychiatry, University Medical Center Groningen, Groningen, The Netherlands

H. van Engeland · M. V. de Jonge
Department of Child and Adolescent Psychiatry and Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands

J. K. Buitelaar · R.-J. van der Gaag
Department of Psychiatry, University Medical Center St. Radboud, Nijmegen, The Netherlands

J. K. Buitelaar · R.-J. van der Gaag
Nijmegen Center for Evidence-Based Practice (NCEBP), Nijmegen, The Netherlands

Recently, Gotham et al. (2009) published calibrated severity scores for the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 1999). Developing these calibrated severity scores was inspired by the need in clinical practice and research for describing the severity of the behavior of children with autism spectrum disorders (ASDs) referring to the core symptoms in the autism spectrum. The ADOS, as a well developed, valid instrument is widely used as an important part of the diagnostic procedure when investigating ASDs in clinical practice and research. The raw scores obtained by an ADOS administration are often used as a means to indicate severity of ASD, even though the ADOS was not specifically developed to facilitate comparison of data from different modules, different moments of measurement or different children. Due to developmental differences between children administered the various modules, to the developmental grading of the items across modules, and to differences between the numbers of items across modules raw scores on the ADOS are not directly comparable.

Compared to the original algorithms (Lord et al. 1999), the revised algorithms (Gotham et al. 2007) increased comparability between modules, including the same item number and the same content per module, although still with different developmental levels or requirements. Additionally, with the revised algorithms the influence of

chronological age and verbal IQ on ADOS outcome decreased for modules 2 and 3 (Gotham et al. 2007, 2008). In independent samples, the validity of the revised algorithms was corroborated, concluding that the sensitivity and specificity were better balanced (modules 2 and 3; De Bildt et al. 2009) and the diagnostic validity improved (modules 1 and 2; Oosterling et al. 2010a).

By calibrating the scores on the revised algorithms, Gotham et al. (2009) intended to further decrease the influence of participant characteristics (e.g. developmental and age differences between children administered different modules) and aimed to "*approximate a severity measure for the construct of 'autism spectrum' as it is measured on the ADOS*" (Gotham et al. 2009, p. 694). In clinical practice and research, the calibrated severity measure should be valuable for comparing ADOS assessments across time and modules; for indicating the severity of specific ASD behavior; for investigating the relationship between severity in ASD and levels of cognitive functioning; for identifying different trajectories in autism severity unrelated to verbal IQ; for describing the behavioral phenotype of ASD well over time and samples; and for selecting more homogeneous groups for studies.

The calibrated severity scores were based on the raw total scores on the revised algorithms (Social Affect and Restricted Repetitive Behaviors; SARRB; Gotham et al. 2007) in 1807 assessments of 1118 individuals with clinical ASD diagnoses. These were divided into eighteen calibration cells, based on the distributions of SARRB scores per age group in the developmental cells corresponding to the revised algorithm groups (cells shown in Fig. 1). Within these 18 cells, raw SARRB scores were converted to a 10-point severity metric, based on percentiles associated with each severity point, with standard scores 1–3 representing the nonspectrum ADOS classification, scores 4–5 ASD classifications and scores 6–10 autism classifications (AD; see Gotham et al. 2009, p. 699 for calibrated severity scores and corresponding raw scores and ADOS classifications).

The severity score distributions were found to be more uniform over the various calibration cells than the ADOS raw score distributions. The distributions of the calibrated severity scores in the separate diagnostic classification groups (AD vs non-autism ASD; non-autism ASD vs nonspectum) were better separated from each other. Additionally, within the ASD classifications the influence of verbal IQ on ADOS outcome decreased from a large to medium effect size. Also, within the total sample, calibrated severity scores were less affected by other participant characteristics (such as age, or maternal education) than raw scores. The authors concluded that the calibrated severity scores are a better means to indicate autism severity than raw ADOS totals, relatively independent from verbal ability.

In the current paper we aim to investigate the validity of the calibrated severity scores in an independent sample of Dutch children administered with ADOS modules 1–3. Based on 1455 assessments from 1248 individuals, our goals are to examine the findings of Gotham et al. (2009): (a) a more uniform distribution of the calibrated severity scores per calibration cell, compared to raw scores; (b) a clearer distinction between the diagnostic groups with respect to the severity of symptoms as measured with the ADOS; and (c) the relationship between the calibrated severity scores and age, level of verbal and non-verbal cognitive functioning, and maternal education. The findings from the Dutch sample will be compared to the original results from Gotham et al. (2009).

Methods

Participants

With data of 1248 individuals the calibrated severity scores on the ADOS as developed by Gotham et al. (2009) were evaluated for modules 1, 2 and 3, for almost all calibration cells except module 2, younger than 5, age 2 (cell 10 in Table 2) and module 3, age 2–5 (cell 16 in Table 2). We were not able to investigate the calibrated severity scores in these groups since the cells did not fulfill our requirement of $n > 15$.

Of these individuals, 107 had two ADOS assessments with contemporaneous clinical classification, and 50 had three ADOS assessments and clinical classifications, resulting in 1455 assessments. All repeated assessments took place within or between modules 1 and 2.

In 542 cases (37.3% of the assessments) the clinical diagnosis was AD, in 486 (33.4%) non-autism ASD (including Asperger Syndrome (AS; 11.3% of non-autism

Fig. 1 N's per age/language level calibration cells. Note: cells only include assessments with clinical ASD classifications; no data available for module 2, age 2 and module 3, age 2–5

Age (in years)		2	3	4	5	6	7	8	9	10+
Module 1	No Words	n=50	n=18	n=28		n=29				
	Single Words	n=88	n=90	n=46	n=37		n=30			
Module 2	Phrases	-	n=33	n=86	n=101		n=29		n=28	
Module 3	Fluent	-				n=176				n=159

ASD), Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS; 88.1%), Rett syndrome (.4%) and Childhood Desintegrative Disorder (CDD; .2%) and in 427 cases nonspectrum (including Mental retardation (MR; 26.2% of nonspectrum), Attention Deficit Hyperactivity Disorder/Oppositional Defiant Disorder (19.7%), language disorder (10.8%) and Anxiety/mood disorder (6.8%), another psychiatric diagnosis (14.3%) and no psychiatric diagnosis (16.9%)).

Data were provided by three Dutch University Centers for Child and Adolescent Psychiatry: in Groningen ($n = 441$), Utrecht ($n = 407$), and Nijmegen ($n = 607$). The ADOS administrations in Groningen en Utrecht had taken place as part of two large studies in the Netherlands concerning the genetics of ASDs. These studies included children referred for child-psychiatric problems/ASD, children recruited for a multi-incidence genetic study and children from an epidemiological study of ASD in mental retardation (population based; De Bildt et al. 2005). This means that not all, and especially not all low-functioning participants from the current study were referred for problems in the autism spectrum, yet they were all evaluated by experienced clinicians. All children from Nijmegen were clinically referred, most of them within the context of an extensive early screening project for ASD (see Oosterling et al. 2010b for more detail).

The ages ranged from 2 through 16 years, with the same age range per module as in the study of Gotham and colleagues (2009; see also Table 1 for a description of the sample).

Differences between our sample and Gotham's sample were tested based on the 95% confidence intervals of the two samples. Significant differences are reported per module, developmental cell and diagnostic group.

Differences Module 1

In module 1 Some Words, all diagnostic groups had higher verbal IQ's in our sample and received lower scores on the ADOS Restricted Repetitive Behavior domain (RRB). Additionally, our autism sample had lower scores on the Autism Diagnostic Interview-Revised (ADI-R) RRB domain, and our nonspectrum group was older. In module 1 No Words, our non-autism ASD sample was older and verbally more capable (verbal IQ measured in only $n = 19$), and our autism sample was lower functioning nonverbally and received lower scores on both ADOS domains.

Differences Module 2

Our module 2, 5 and older Autism sample had a higher verbal IQ. This same group also scored significantly lower on the ADOS SA and RRB domains, and ADI-R Social and Verbal Communication domains. The only difference between Gotham's and our sample of module 2, 5 and older non-autism ASD yielded the lower scores of our group on the RRB domain of the ADOS. Our nonspectrum group of module 2, 5 and older had lower scores than Gotham's group on the RRB domains of ADI-R and ADOS, and on the Verbal Communication domain of the ADI-R.

For module 2, younger than 5, all our groups were older. Additionally, our autism group had lower scores on the ADI-R domains Verbal Communication and RRB, and on both ADOS domains. Our non-autism ASD group was verbally more capable and received lower scores on both ADOS domains. Our nonspectrum group had lower scores on the RRB domain of the ADI-R and ADOS, compared to Gotham's sample.

Differences Module 3

For module 3, our sample was older, had lower scores on the RRB domains of ADI-R and ADOS. Our nonspectrum group was lower functioning. Our non-autism ASD group had higher scores on Non-verbal Communication of the ADI-R.

Instruments

The ADOS and ADI-R were administered by trained psychologists or psychiatrists who met standard requirements of reliability and administration in research. The Autism Diagnostic Interview-Revised (ADI-R; Rutter et al. 2003) was available for 1038 assessments (71.3% of the sample).

Clinical classifications were based on DSM-IV-TR (APA 2000) criteria, and assigned by experienced clinicians, reviewing all available diagnostic information. The far majority of clinical classifications was established by a multidisciplinary team comprising minimally a child psychiatrist and a psychologist. The research classifications in one of the genetic studies (Groningen) were based on reviewing all available diagnostic information by one experienced clinician ($n = 265$; see also De Bildt et al. 2005).

Measures of cognitive functioning were available for 1086 assessments (74.6% of the sample), based on standardized tests. In module 1, the majority of cases was tested with the Mullen Scales of Early Learning (MSEL; Mullen 1995), the Psycho Educational Profile-Revised (PEP-R; Schopler et al. 1990), or a Dutch nonverbal intelligence test, the Snijders-Oomen Niet-verbale intelligentie test-Revisie (SON-R; Snijders et al. 1996). In module 2, the SON-R and WPPSI-R (Wechsler 1989; Vander Steene and Bos 1997) were administered most frequently, and in module 3 Wechsler Scales (WISC-III-

Table 1 Sample description

DX	Module 1, no words				Module 1, some words				Module 2, younger than 5				Module 2, 5 and older				Module 3			
	N	Mean	SD	Range	*N*	Mean	SD	Range	*N*	Mean	SD	Range	*N*	Mean	SD	Range	*N*	Mean	SD	Range
Autism																				
Age	112	54.04	32.0	24–178	209	54.82	28.73	24–177	54	51.73	6.15	39–60	68	95.22	35.83	61–189	99	121.14	31.14	73–190
viq	15	28.22	16.85	9–65	49	63.12	21.63	6–102	9	89.44	27.91	44–138	39	70.97	24.43	17–128	83	91.46	23.49	24–136
nviq	28	36.23	21.47	5–89	123	66.35	23.83	5–113	42	88.17	21.75	50–125	56	80.88	25.79	16–137	89	92.90	24.05	17–155
PEP-R Ratio IQ	61	39.56	13.05	11–71	48	52.44	13.10	28–84	–	–	–	–	–	–	–	–	–	–	–	–
ADI social	88	19.66	5.83	4–29	141	18.23	5.98	5–30	29	15.52	5.71	5–27	58	19.33	6.65	4–30	96	20.21	5.69	8–29
ADI comm-V	9	15.67	7.58	9–32	70	14.23	4.52	2–25	29	11.31	5.33	3–23	58	14.86	5.03	4–25	96	16.31	4.76	3–25
ADI comm-NV	88	11.36	2.78	3–14	139	9.80	3.17	0–14	28	6.46	3.94	0–14	57	9.14	3.66	0–14	95	10.13	3.33	2–14
ADI-RR	88	4.51	2.17	0–10	141	4.55	2.63	0–11	29	4.55	2.73	0–10	58	6.14	2.85	0–12	96	5.96	2.68	1–12
ADOS SA	112	16.16	2.37	4–18	209	14.01	4.72	0–20	54	10.61	4.22	2–18	68	10.59	4.87	0–19	99	10.99	4.51	0–19
ADOS RR	112	3.63	1.60	0–7	209	2.60	1.73	0–7	54	1.98	1.88	0–8	68	2.29	1.66	0–7	99	1.87	1.54	0–6
Non-autism ASD																				
Age	13	80.10	50.98	13–24	82	47.71	30.18	26–179	66	51.75	5.66	37–60	89	84.41	28.78	61–201	236	122.58	28.79	73–202
viq	2	55.83	3.54	53–58	42	85.17	21.03	50–178	14	102.65	9.92	91–126	31	71.19	28.84	23–124	184	97.40	23.47	34–148
nviq	7	46.43	39.95	10–100	57	82.04	26.13	16–133	49	99.39	19.12	58–148	69	78.80	27.57	16–130	207	96.10	22.62	27–155
PEP-R ratio IQ	3	47.33	10.26	36–56	11	49.64	15.17	30–77	–	–	–	–	–	–	–	–	–	–	–	–
ADI social	9	17.22	7.29	4–26	64	11.47	6.33	0–26	27	13.74	7.77	2–29	59	14.37	6.35	1–27	196	16.42	6.49	0–29
ADI comm-V	0	–	–	–	44	10.52	5.08	0–19	27	9.89	5.84	0–20	58	10.50	4.88	2–21	196	12.91	5.14	0–26
ADI comm-NV	9	9.78	3.90	5–14	62	7.34	3.92	0–14	26	6.27	4.02	0–14	57	6.39	3.96	0–14	195	7.95	3.81	0–14
ADI-RR	9	3.11	2.37	0–7	64	3.45	2.34	0–9	27	4.33	3.11	0–11	59	3.78	3.00	0–12	196	4.24	2.57	0–12
ADOS SA	13	13.85	3.46	6–18	82	9.22	4.35	1–19	66	6.27	3.61	1–13	89	7.42	4.01	0–18	236	8.00	4.52	0–19
ADOS RR	13	2.69	1.89	0–6	82	1.50	1.50	0–6	66	1.24	1.39	0–5	89	1.38	1.52	0–6	236	1.14	1.35	0–8
Nonspectrum																				
Age	10	73.99	49.05	28–142	119	51.92	34.37	25–165	74	49.26	7.16	28–60	114	96.19	38.21	61–201	110	126.44	28.06	67–200
viq	2	70.36	21.72	55–86	73	85.03	19.45	18–134	12	99.02	17.32	73–129	48	56.10	24.90	22–111	82	76.82	28.14	24–135
nviq	2	55.11	60.96	12–98	105	80.26	29.19	10–140	52	101.48	17.02	66–133	91	71.22	28.73	16–133	86	77.06	26.90	33–130
PEP-R Ratio IQ	3	68.33	8.08	61–77	2	56.50	9.19	50–63	–	–	–	–	–	–	–	–	–	–	–	–
ADI social	8	12.13	7.86	1–25	92	7.77	5.22	0–28	14	7.79	4.92	1–18	76	10.71	6.99	0–27	81	10.06	7.25	0–26
ADI comm-V	0	–	–	–	62	6.48	3.85	1–22	13	5.62	4.03	1–13	73	7.19	4.57	0–20	81	7.85	5.57	0–20
ADI comm-NV	8	7.75	4.50	3–14	92	4.97	3.35	0–14	13	3.46	2.26	1–9	76	4.66	3.58	0–14	80	5.34	4.28	0–14
ADI-RR	8	2.75	2.05	1–7	92	2.27	1.88	0–8	14	1.64	1.39	0–5	76	2.57	2.38	0–12	81	2.40	2.55	0–9
ADOS SA	10	7.30	5.08	2–16	119	4.52	3.81	0–19	74	3.05	2.76	0–12	114	4.73	3.60	0–16	110	4.56	4.09	0–17
ADOS RR	10	2.00	1.76	0–5	119	.69	1.01	0–4	74	.49	.82	0–3	114	.94	1.17	0–6	110	.55	.91	0–4

age age in months, *viq* verbal IQ, *nviq* non-verbal IQ, *PEP-R Ratio IQ* (PEP-R develomental age/chronological age) * 100, *ADI social* ADI-R reciprocal social interaction total, *ADI comm-V* ADI-R communication total for verbal children, *ADI comm-NV* ADI-R communication total for non-verbal children, *ADI RR* ADI-R restricted, repetitive behaviors total, *ADOS SA* ADOS social affect (revised algorithm), *ADOS RR* ADOS restricted, repetitive behaviors (revised algorithm)

NL (Wechsler 1992; Kort et al. 2005), WISC-R (Wechsler 1974; Vander Steene et al. 1986), WPPSI-R), or RAVEN progressive matrices (Raven 1995, 1996) were most frequently applied.

Design and Analysis

First, differences in reported raw and calibrated severity scores between the study of Gotham and colleagues and the current study were tested based on 95% confidence interval calculations using mean scores, SD and sample size. When the confidence intervals (partially) overlap, the difference between the reported mean scores is not significant. A found difference is only significant ($p < .05$) when the 95% confidence intervals of the mean scores do not show any overlap. Second, to examine the expected increase in uniformity of the distributions of the calibrated severity scores, these distributions were obtained for each age/language cell (Fig. 1), and compared to the distributions of raw scores per age/language cell, for participants with a clinical ASD classification only (in accordance with Gotham et al. 2009). Third, the distributions of calibrated severity scores were compared over the three diagnostic groups (AD, non-autism ASD, nonspectrum) in order to investigate the expected increase in heterogeneity of the severity distribution per diagnostic group. Last, to investigate the relationship between the raw and calibrated severity scores on one hand and age, level of verbal and non-verbal cognitive functioning and maternal education on the other, Pearson r correlations were computed.

Results

Comparison of the Raw and Calibrated Severity Scores to the Findings of Gotham et al. (2009)

In the current sample, participants with a clinical ASD classification showed consistently lower raw scores compared to the sample of Gotham et al., on the ADOS SARRB algorithm for all modules, yet most so for module1 No Words and both modules 2. Mean scores of the current sample are shown in Table 2, with mean scores from Gotham's sample in parentheses. Significantly lower raw scores than in Gotham's sample (2009) were reported for ten out of 16 cells investigated in this study.

Raw score differences from Gotham's sample varied from one point to six (mod 2, ages 9–16) or seven points (m2 ages 5–6 and m2 ages 7–8). As can be seen in Table 1, the current sample had low RRB-scores in module 1 Some Words and module 3 (significantly lower than Gotham's sample (2009)). In these modules, the differences seem to be caused to a large extent by the lower RRB scores. In module 2 (younger than 5 and 5 and older) and module 1 No Words, there was a difference in the clinical ASD groups in SARRB scores (and not RRB only). The ADOS-classification based on the mean calibrated severity scores was AD for eight out of 16 cells, identical to the classifications in these cells in the study of Gotham et al. (2009) The other eight fell into the ASD range, differing from Gotham's group (all AD classifications). The comparisons additionally showed that, with respect to the height of the scores, the assessments of the ADOS in ten of the 16 cells in the Dutch sample significantly differed from those in Gotham's sample. In seven out of these ten cells, this lead to a lower classification. In three cells the lower scores did not result in lower classifications. However in one cell (module 3, age 6–9) scores did not differ significantly, yet lead to a lower classification in the Dutch sample.

Distributions of Raw and Calibrated Severity Scores Over Age/Language Level for ASD Classifications

The current sample showed distributions of raw scores (shown in Fig. 2), comparable to Gotham et al. (2009), although lower, in module 1 No Words, module 1 Some Words and module 3. Cells 11–15 (module 2 groups) showed a different pattern: the raw score distributions in the current sample were lower and did not show a gradual increase towards higher scores for older age groups.

For module 1, the distributions of calibrated severity scores per calibration cell (Fig. 3) were more uniform than the distributions of raw scores. For module 2, the severity scores were rather uniform, yet low (in the lower range of or below the AD-range) and not clearly more uniform than the raw score distributions. The calibrated severity score distributions in module 3 were broad, ranging from 3 through 8 or 9. The calibrated severity score distributions were not clearly more uniform than the raw score distributions.

Distributions of Raw and Calibrated Severity Scores Over Diagnostic Groups

The distributions of calibrated severity scores differed between the three diagnostic groups (clinical classifications of AD, non-autism ASD and nonspectrum; Fig. 5), except for the clinical non-autism ASD group. The scores in this group overlapped with the AD group on one point, with scores in the AD range (calibrated severity score of 6) and slightly with the nonspectrum group (with scores on the borderline between nonspectrum and non-autism ASD range). Calibrated severity scores were not better separated from each other between diagnostic groups than raw scores (see for raw scores Fig. 4).

Table 2 Raw score and calibrated severity score means and standard deviations by calibration cell (assessments with clinical ASD classifications only)

Cell	Module and age group	SARRB algorithm raw total score			Calibrated severity scores		
		N	*M*	SD	*M*	SD	ADOS classification
1	Mod 1, NW, age 2	50	19.12 (20.13)	3.93 (4.83)	6.72 (7.29)	1.67 (2.11)	AUT
2	Mod 1, NW, age 3	18	18.44 (21.63)*	4.27 (3.85)	5.94 (7.56)*	1.59 (1.85)	ASD
3	Mod 1, NW, age 4–5	28	20.32 (21.96)*	2.41 (3.63)	7.11 (7.87)*	.99 (1.48)	AUT
4	Mod 1, NW, age 6–14	29	19.79 (22.35)*	2.83 (3.34)	6.62 (7.88)*	.98 (1.45)	AUT
5	Mod 1, SW, age 2	88	12.76 (15.64)*	6.27 (5.77)	5.75 (7.02)*	2.74 (2.45)	ASD
6	Mod 1, SW, age 3	90	14.42 (15.85)	6.13 (5.37)	6.32 (6.99)	2.61 (2.26)	AUT
7	Mod 1, SW, age 4	46	15.07 (17.13)	4.72 (5.95)	6.52 (7.21)	1.76 (2.16)	AUT
8	Mod 1, SW, age 5–6	37	17.30 (18.84)	4.70 (4.71)	6.97 (7.48)	1.62 (1.72)	AUT
9	Mod 1, SW, age 7–14	30	19.93 (20.68)	4.14 (4.24)	7.57 (7.97)	1.85 (1.77)	AUT
10	Mod 2, phrases, age 2	–	–	–	–	–	–
11	Mod 2, phrases, age 3	33	10.96 (14.57)*	5.62 (5.01)	5.70 (7.38)*	2.70 (2.04)	ASD
12	Mod 2, phrases, age 4	86	9.36 (14.43)*	5.04 (5.93)	4.65 (6.73)*	2.40 (2.44)	ASD
13	Mod 2, phrases, age 5–6	101	9.61 (16.84)*	5.08 (5.78)	4.71 (7.45)*	2.24 (1.99)	ASD
14	Mod 2, phrases, age 7–8	29	11.07 (18.49)*	5.18 (5.22)	5.28 (7.79)*	2.09 (1.71)	ASD
15	Mod 2, phrases, age 9–16	28	13.46 (19.16)*	5.71 (4.48)	6.07 (8.10)*	2.21 (1.37)	AUT
16	Mod 3, fluent, age 2–5	–	–	–	–	–	–
17	Mod 3, fluent, age 6–9	176	10.50 (11.66)	5.53 (5.19)	5.98 (6.64)	2.83 (2.55)	ASD
18	Mod 3, fluent, age 10–16	159	9.94 (12.48)*	5.11 (4.94)	5.75 (7.09)*	2.75 (2.45)	ASD

ADOS classification based on mean calibrated severity scores of each calibration cell. Scores from Dutch sample with scores from Gotham's sample in parentheses. ADOS Classification in Dutch sample, when different from Gotham's sample: ADOS classification in Gotham's sample in parentheses. * = scores significantly lower in the Dutch sample compared to the sample of Gotham et al. (2009), based on 95% Confidence Intervals. For cells 10 and 16, no Dutch data are available

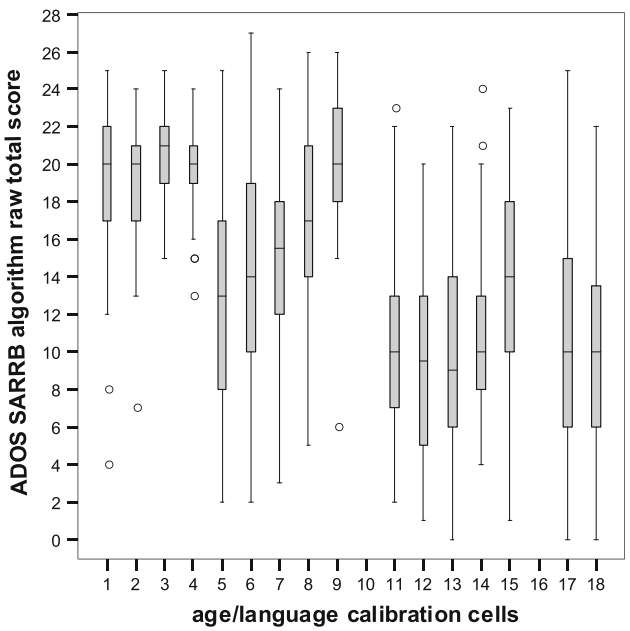

Fig. 2 Distributions of ADOS SARRB raw total scores per cell (assessments with clinical ASD classifications only). Note: See Table 2 for a key of the 18 cells

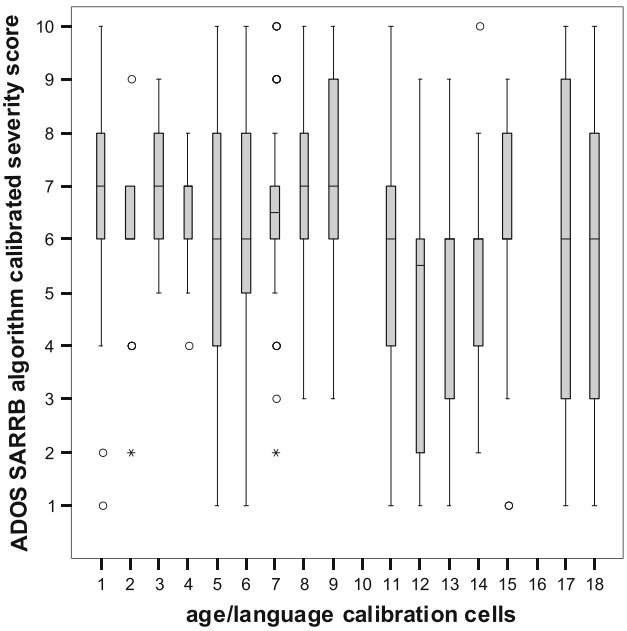

Fig. 3 Distributions of ADOS SARRB calibrated severity scores per cell (assessments with clinical ASD classifications only). Note: See Table 2 for a key of the 18 cells

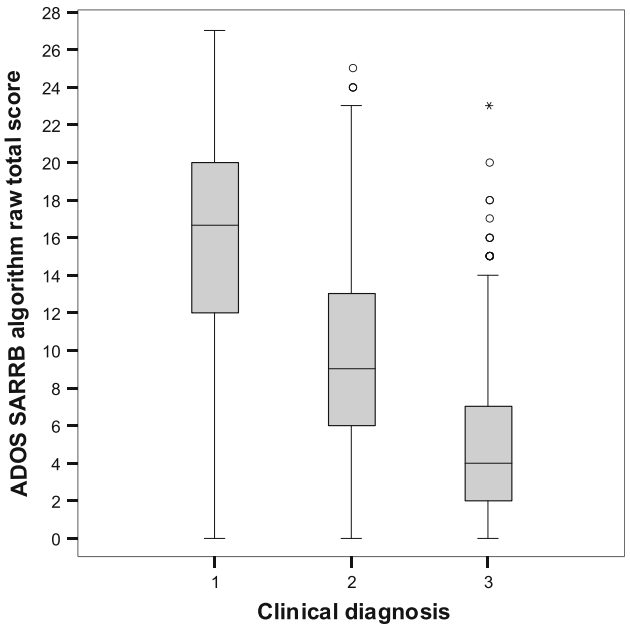

Fig. 4 Distributions of ADOS SARRB raw scores over clinical classifications. Note: *1* clinical AD classification (n = 545); *2* clinical non-autism ASD classification n = 491); *3* clinical nonspectrum classification (n = 427)

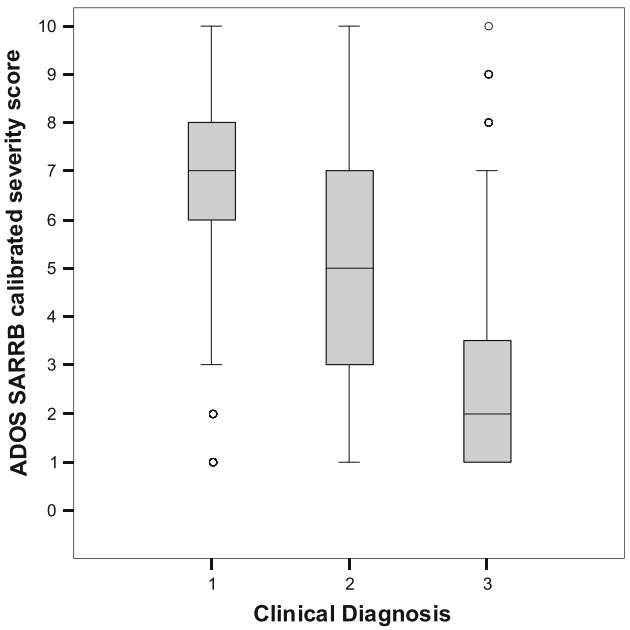

Fig. 5 Distributions of ADOS SARRB calibrated severity scores over clinical classifications. Note: *1* clinical AD classification (n = 542); *2* clinical non-autism ASD classification n = 486); *3* clinical nonspectrum classification (n = 427)

With respect to maternal education and age, no meaningful correlations were found with either the raw scores on the SARRB algorithm (Pearson r for maternal education .029, $p = .433$, $n = 727$; for age $r = -.064$, $p < .001$, $n = 1455$) or the calibrated severity scores (Pearson r for maternal education .028, $p = .447$, $n = 727$; for age $r = .069$, $p = .008$, $n = 1455$). Levels of verbal and non-verbal cognitive functioning showed a higher correlation with raw scores on the SARRB algorithm (Pearson r for PEP Ratio IQ $-.398$ $p < .001$, $n = 130$; for VIQ $r = -.252$ $p < .001$, $n = 685$; for NVIQ $r = -.333$ $p < .001$, $n = 1063$) than with calibrated severity scores on the SARRB algorithm (Pearson r for PEP Ratio IQ $-.235$ $p = .007$, $n = 130$; for VIQ $r = -.128$ $p = .001$, $n = 685$; for NVIQ $r = -.209$ $p < .001$, $n = 1063$). Although all values except for maternal education were significant, none of them were high enough to be of any meaning.

Discussion

The results from the current study corroborate the calibrated severity scores of the ADOS as developed by Gotham et al. (2009) for module 1 and to a lesser extent module 3, in an independent sample. The calibrated severity scores discriminate the clinical AD, non-autism ASD and nonspectrum classifications well, and are more comparable over various developmental cells than the raw scores on the ADOS, especially in module 1, and somewhat less so in module 3. The calibrated severity scores show good validity in this independent sample, even when considering the differences between the current and Gotham's sample. The Dutch sample with a clinical AD classification contained children with relatively low scores on the ADOS and ADI-R, indicating that this group was not on the severest end of the autism spectrum. Still, for module 1 and to a lesser extent for module 3 the calibrated severity scores were replicated. Additionally, the relatively high ages (in the non-autism ASD groups) did not affect the validity of the calibrated severity scores in modules 1 and 3. Differences in levels of verbal and non-verbal cognitive functioning between the current and Gotham's sample did not affect the validity either. In modules 1 and 3, the ADOS algorithm showed to be rather independent from various predicting factors. The relationship between the ADOS algorithm raw scores and age and level of maternal education was already small, and did not change when computed for the calibrated severity scores. Calibrated severity scores showed to be less related to level of cognitive functioning than raw scores, indicating a greater independence from this factor. Therefore, the current study corroborates that the calibrated severity scores are a valid measure for severity of autism spectrum features with respect to module 1 and module 3.

For module 2, the current study does not replicate the metric of calibrated severity scores. Although the

calibrated severity scores were more independent from level of cognitive functioning, their distributions were not more uniform than the raw score distributions. The mean calibrated severity scores were low, which is most probably due to the characteristics of the Dutch module 2 sample, including children with low scores on the ADI-R and especially ADOS in the AD and non-autism ASD groups. The low raw scores inevitably lead to lower calibrated severity scores. The question is why children with a clinical A(S)D diagnosis have received such low scores on the ADOS module 2. Age does not seem to be of any influence, since the Dutch children administered with module 2 were only 1 to 9 months older than in Gotham's group, comparable to the age difference between the samples of modules 1 and 3. However, a higher age in combination with higher verbal IQ's may indicate a choice for a too easy module for some children administered with module 2. Consequently, as reported by Klein-Tasman et al. (2007), children with AD may have been identified, whereas children with non-autism ASD may have reached ADOS-scores outside the autism spectrum. Another explanation for lack of replication for module 2 and a less clear replication for module 3 may be that the proportions of children with a clinical diagnosis of AD versus non-autism ASD differ enormously between the Dutch sample and the sample of Gotham et al. (2009). In Gotham's study, in module 2 children with AD outnumbered children with non-autism ASD (253 (61.4%):159), in the Dutch sample this was the other way around (125 (44.5%):156). For module 3, Gotham studied 178 (41.6%) children with AD and 250 with non-autism ASD, whereas the Dutch module 3 sample contained 99 (29.2%) children with AD and 240 with non-autism ASD. This will very likely have influenced the results, due to lower scores on the ADOS for children with non-autism ASD compared to children with AD and perhaps more irregular scoring patterns on the ADOS in children with a final non-autism ASD diagnosis.

Another remarkable finding from the current study is that for all modules the Dutch sample showed significantly lower scores on the ADOS RRB domain, except for module 1 No Words non-autism ASD and nonspectrum group. For one reason or another, RRB is less frequently reported in the Dutch sample. This finding does not seem to be related to the ADOS administration as such, since most groups showed lower RRB scores on the ADI-R as well. One explanation may be that groups from the current sample were often older than Gotham's sample. Reports of abating repetitive restricted behavior with age (see for example Esbensen et al. 2009) indicate that RRBs may tend to occur less often in the Dutch sample than in Gotham's sample, simply because of age. However, first this does not take into account the lower RRB scores on the ADI-R, which reflects such behaviors during development based on their 'ever' scores. Second, the age difference is not that large that this explanation would be very likely. It is not plausible either that the level of cognitive functioning influences this finding. Not only the A(S)D groups showed lower RRB scores, yet also the nonspectrum groups had lower scores RRB than Gotham's nonspectrum groups. Whether the difference in reported RRB is a difference in actual occurrence of the behaviors or in the identification of these behaviors as RRB is unclear. To our knowledge, intercultural differences with respect to RRB have not been reported. This issue cannot be further investigated with the current data. More research on this difference is nevertheless very important, especially with the current development towards the DSM-V, in which the RRB domain will be more prominently present.

Limitations

As was the case in Gotham's study, the current sample was not a population based sample. Additionally, the fact that the ADOS administrations that were analyzed were collected over a 10 year period may have influenced the results, due to the changes in identification of ASDs. Last, the current study does not elaborate the knowledge on longitudinal variations or developmental trajectories within children. Due to the small amount of repeated ADOS administrations, and more so due to the small range in which these repeated measures took place (within or between modules 1 and 2) trajectories could not be well investigated. More important, numbers of participants in some calibration cells were so small (e.g. cell 10 and cell 16, Table 2), that replication of the metric of calibrated severity scores could not be performed for children in module 2, age 2 and module 3 age 2–5.

Conclusion

The greater comparability between ADOS administrations at different times, ages and in different modules, as reached by Gotham et al. (2009) with the calibrated severity measures, seems to be corroborated by the current study for module 1 and to a lesser extent for module 3. This replication endorses the value of the calibrated severity scores as a way to compare ADOS scores across time, age and module and its value as a measure of ASD severity. For module 2, the calibrated severity scores need to be further investigated within a sample that resembles Gotham's sample in age and level of verbal functioning. The fact that the results were replicated in module 1 is especially promising, since nowadays ASDs are identified earlier, and calibrated severity measures seem to be an indication of

severity of symptoms, and adaptive and problem behavior, outcome and changes over time (Gotham et al. 2009). The earlier an ASD is identified, and the better the clinical characteristics and severity of the symptoms can be specified at that time, the better and more specific care can be provided, and the better its effect can be measured over time. However, it should be kept in mind that the ADOS, and therefore its calibrated severity scores, are only part of the diagnosis of an individual with an ASD, which should be extended with broader information from various sources in order to complete the clinical picture.

Acknowledgments This research was supported by the Korczak Foundation and ZON-MW.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, 4th edition, text revision (DSM-IV-TR)*. Washington, DC: American Psychiatric Association.

De Bildt, A., Sytema, S., Kraijer, D., & Minderaa, R. (2005). Prevalence of pervasive developmental disorders in children and adolescents with mental retardation. *Journal of Child Psychology and Psychiatry, 46*, 275–286.

De Bildt, A., Sytema, S., Van Lang, N. D. J., Minderaa, R. B., Van Engeland, H., & De Jonge, M. V. (2009). Evaluation of the ADOS revised algorithm: The applicability in 558 Dutch children and adolescents. *Journal of Autism and Developmental Disorders, 39*, 1350–1358.

Esbensen, A. J., Seltzer, M. M., Lam, K. S. L., & Bodfish, J. W. (2009). Age-related differences in restricted repetitive behaviors in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*, 57–66.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*, 693–705.

Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., et al. (2008). A replication of the autism diagnostic observation schedule (ADOS) revised algorithms. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 642–651.

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders, 37*, 613–627.

Klein-Tasman, B. P., Risi, S., & Lord, C. E. (2007). Effect of language and task demands on the diagnostic effectiveness of the autism diagnostic observation schedule: The impact of module choice. *Journal of Autism and Developmental Disorders, 37*, 1224–1234.

Kort, W., Schittekatte, M., Bosmans, M., Compaan, E. L., Dekker, P. H., Vermeir, G., et al. (2005). *Wechsler intelligence scale for children-III. Nederlandstalige Uitgave*. Amsterdam: Pearson.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *ADOS. Autism diagnostic observation schedule. Manual*. Los Angeles: WPS.

Mullen, E. (1995). *Mullen scales of early learning. AGS Edition*. Circle Pines, MN: American Guidance Service.

Oosterling, I., Roos, S., de Bildt, A., Rommelse, N., de Jonge, M., Visser, J et al. (2010a) Improved diagnostic validity of the ADOS revised algorithms: A replication study in an independent sample. *Journal of Autism and Developmental Disorders, 40*(6), 689–703.

Oosterling, I. J., Wensing, M., Swinkels, S. H., Gaag, R. J., van der Visser, J. C., Woudenberg, T., et al. (2010b). Advancing early detection of autism spectrum disorder by applying a two-stage screening approach. *Journal of Child Psychology and Psychiatry, 3*(51), 250–258.

Raven, J. C. (1995). *Colored progressive matrices sets I and II* (1996th ed.). Oxford: Oxford Psychologists Press Ltd.

Raven, J. C. (1996). *Progressive matrices: A perceptual test of intelligence. Individual form* (1996th ed.). Oxford: Oxford Psychologists Press Ltd.

Rutter, M., Le Couteur, A., & Lord, C. (2003). *ADI-R. Autism diagnostic interview revised. Manual*. Los Angeles: Western Psychological Services.

Schopler, E., Reichler, R. J., Bashford, A., Lansing, M. D., & Marcus, L. M. (1990). *The Psychoeducational Profile Revised (PEP-R)*. Austin: Pro-Ed.

Snijders, J. Th., Tellegen, P. J., Winkel, M., & Laros, J. A. (1996). *SON-R 2, 5–7 Snijders-Oomen Niet-verbale Intelligentietest-Revisie [SON-R 2, 5–7 Snijders-Oomen Non-verbal intelligence test-revised]*. Lisse, the Netherlands: Swets & Zeitlinger.

Vander Steene, G., & Bos, A. (1997). *Wechsler preschool and primary scale of intelligence-revised. Vlaams-Nederlandse Aanpassing*. Lisse, the Netherlands: Swets & Zeitlinger.

Vander Steene, G., Haasen, P. Pv., De Bruyn, E. E. J., Coetsier, P., Pijl, Y. J., Poortinga, Y. H., et al. (1986). *Wechsler intelligence scale for children-revised. Nederlandstalige Uitgave*. Lisse, the Netherlands: Swets & Zeitlinger.

Wechsler, D. (1974). *Wechsler intelligence scale for children-revised*. New York: Psychological Corporation.

Wechsler, D. (1989). *Wechsler preschool and primary scale of intelligence-revised*. New York: Psychological Corporation.

Wechsler, D. (1992). *Wechsler intelligence scale for children-III*. New York: Psychological Corporation.